Speaker Localization and Detection in Videoconferencing Environments Using a Modified SRP-PHAT Algorithm

A. Marti, M. Cobos, E. Aguilera, J. J. Lopez
Instituto de Telecomunicaciones y Aplicaciones Multimedia,
Universitat Politècnica de València,
8G Building - access D - Camino de Vera s/n - 46022 Valencia (Spain)
Corresponding author: ammargue@iteam.upv.es

Abstract

The Steered Response Power - Phase Transform (SRP-PHAT) algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments. However, its practical implementation is usually based on a costly fine grid-search procedure, making the computational cost of the method a real issue. In this paper, we introduce an effective strategy which performs a full exploration of the sampled space rather than computing the SRP at discrete spatial positions, increasing its robustness and allowing for a coarser spatial grid that reduces the computational cost required in a practical implementation. The modified SRP-PHAT functional has been successfully implemented in a real time speaker localization system for multi-participant videoconferencing environments. Moreover, a localization-based speech-non speech frame discriminator is presented.

Keywords: Sound source localization, SRP-PHAT, microphone array, speaker detection, speech enhancement.

1. Introduction

The localization of emitting signal sources has been the focus of attention for more than a century. Arrays of microphones have a variety of applications in speech data acquisition systems. Applications include teleconferencing, biomedical devices for hearing impaired persons, audio surveillance, gunshot detection, speech enhancement and camera pointing systems. The fundamental requirement for sensor array systems is the ability to locate and track a signal source. In addition to having high accuracy, the location estimator must be capable of a high update rate at reasonable computational cost in order to be useful for real time tracking and beamforming applications. Source location data may also be used for purposes other than beamforming, such as aiming a camera in a video conferencing system [1, 2].

Many current sound source localization (SSL) systems assume that the sound sources are distributed on a horizontal plane [3]. This assumption simplifies the problem of SSL in almost all methods. For example, in teleconference applications they assume all talkers speak at the same height which is somewhat true. Moreover, in most dominant SSL methods, the computational cost for two dimensional cases is high so that the real time implementation needs a computer with high processing power.

Algorithms for SSL can be broadly divided into indirect and direct approaches [4]. Indirect approaches usually follow a two-step procedure: they first estimate the Time Difference Of Arrival

(TDOA) [5] between microphone pairs and, afterwards, they estimate the source position based on the geometry of the array and the estimated delays. On the other hand, direct approaches perform TDOA estimation and source localization in one single step by scanning a set of candidate source locations and selecting the most likely position as an estimate of the source location. In addition, information theoretic approaches have also shown to be significantly powerful in source localization tasks [6]. The SRP-PHAT algorithm is a direct approach that has been shown to be very robust under difficult acoustic conditions [7, 8, 9]. The algorithm is commonly interpreted as a beamforming-based approach that searches for the candidate source position that maximizes the output of a steered delay-and-sum beamformer. However, despite its robustness, computational cost is a real issue because the SRP space to be searched has many local extrema [10].

Very interesting modifications and optimizations have already been proposed to deal with this problem, such as those based on Stochastic Region Contraction (SRC) [11] and coarse-to-fine region contraction [12], achieving a reduction in computational cost of more than three orders of magnitude. In this paper, we propose a different strategy where, instead of evaluating the SRP functional at discrete positions of a spatial grid, it is accumulated over the Generalized Cross Correlation (GCC) lag space corresponding to the volume surrounding each point of the grid. The benefits of following this approach are twofold. On the one hand, it incorporates additional spatial knowledge at each point for making a better final decision. On the other hand, the proposed modification achieves the same performance as SRP-PHAT with fewer functional evaluations, relaxing the computational demand required for a practical application.

2. *Srp-Phat* Algortihm

To localize a sound source, a beamformer approach can be followed by scanning (or steering) over a predefined spatial region by adjusting its steering delays. The output of a beamformer, when used in this way, is known as the steered response.

Consider the output from microphone l, $m_l(t)$, in a M microphones system. Then the SRP at the spatial point $\boldsymbol{x} = [x,y,z]$ for a time frame n of length T is defined as

$$P_n(\boldsymbol{x}) \equiv \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l\left(t + \tau(\boldsymbol{x}, l)\right) \right|^2 dt \,,$$

[1]

where w_l is a weight and $\tau(\boldsymbol{x},l)$ is the direct time of travel from location X to microphone l.

Dibiase [10] showed that the SRP can be computed by summing the GCCs for all possible pairs of the set of microphones. The GCC for a microphone pair (k,l) is computed as

$$R_{m_k m_l}(\tau) = \int_{-\infty}^{\infty} \Phi_{kl}(\omega) M_k(\omega) M_l^*(\omega) e^{j\omega\tau}\, d\omega \,,$$

[2]

Where τ is the time lag, * denotes complex conjugation, $M_l(\omega)$ is the Fourier transform of the microphone signal $m_l(t)$, and $\Phi_{kl}(\omega) = W_k(\omega) W_l^*(\omega)$ is a combined weighting function in the frequency domain. The Phase Transform (PHAT) [13] has been demonstrated to be a very effective GCC weighting for time delay estimation in reverberant environments:

$$\Phi_{kl} \equiv \frac{1}{|M_k(\omega) M_l^*(\omega)|} \,.$$

[3]

Taking into account the symmetries involved in the computation of Eq. (1) and removing some fixed energy terms [10], the part of $P_n(\boldsymbol{x})$ that changes with $\boldsymbol{x}$ is isolated as

$$P'_n(\boldsymbol{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} R_{m_k m_l}\left(\tau_{kl}(\boldsymbol{x})\right) \,,$$

[4]

where $\tau_{kl}(\boldsymbol{x})$ is the Inter-Microphone Time Delay Function (IMTDF). This function is very important, since it represents the theoretical direct path delay for the microphones pair (k,l) resulting from a point source located at $\boldsymbol{x}$. The IMTDF is mathematically expressed as

$$\tau_{kl}(\boldsymbol{x}) = \frac{\|\boldsymbol{x} - \boldsymbol{x}_k\| - \|\boldsymbol{x} - \boldsymbol{x}_l\|}{c} \,,$$

[5]

where c is the speed of sound, and x_k and x_l are the microphone locations.

The SRP-PHAT algorithm consists in evaluating the functional $P'_n(\boldsymbol{x})$ on a fine grid G with the aim of finding the point-source location $\boldsymbol{x}_s$ that provides the maximum value:

$$\boldsymbol{x}_s = \arg\max_{\boldsymbol{x} \in G} P'_n(\boldsymbol{x}) \,.$$

[6]

3. Improved Srp-Phat algorithm for source localization

A different strategy for implementing a less cost computational SRP-PHAT algorithm is shown

The SRP-PHAT algorithm is a direct approach that has been shown to be very robust under difficult acoustic conditions.

The advantage of the proposed method relies on the reduced number of required functional evaluations for detecting the true source location.

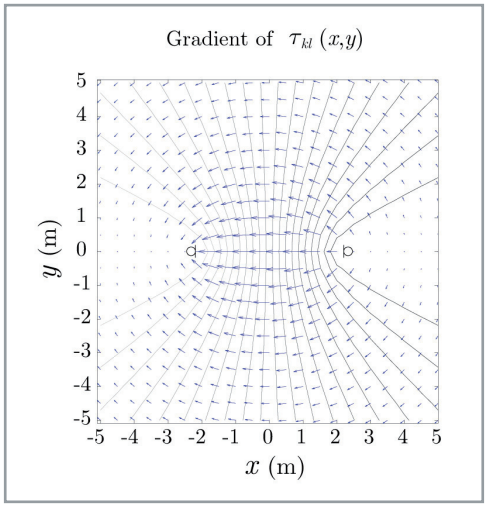

Figure 1. *Example of IMTDF. Gradient.*

in this section. As commented above, the IMTDF plays a very important role in the source localization task. This function can be interpreted as the spatial distribution of possible TDOAs resulting from a given microphone pair geometry.

The function $\tau_{kl}(\boldsymbol{x})$ is continuous in $\boldsymbol{x}$ and changes rapidly at points close to the line connecting both microphone locations. Therefore, a pair of microphones used as a time-delay sensor is maximally sensible to changes produced over this line. The gradient of the function is shown in Figure 1.

It is useful here to remark that the equation $|\tau_{kl}(\boldsymbol{x})| = C$, with C being a positive real constant, defines a hyperboloid in space with foci on the microphone locations $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$. Moreover, the set of continuous confocal half-hyperboloids $\tau_{kl}(\boldsymbol{x}) = C$, with $C \in [-Cmax, Cmax]$, being $Cmax = (1/c)\,||x_k - x_l||$, spans the whole three-dimensional space.

At this point we can formulate the next theorem: Given a volume V in space, the IMTDF for points inside V, $\tau_{kl}(\boldsymbol{x} \in V)$, takes only values in the continuous range [min $(\tau_{kl}(\boldsymbol{x} \in \partial V))$, max $(\tau_{kl}(x \in \partial V))$] where ∂V is the boundary surface that encloses V.

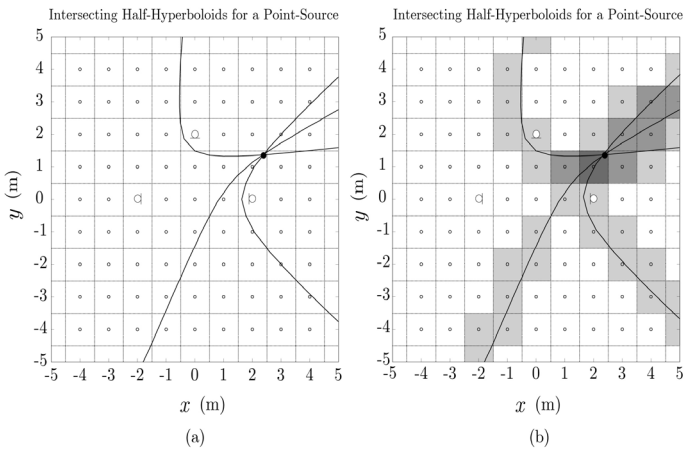

Figure 2. *Intersecting half-hyperboloids and localization approaches. (a) Conventional SRP-PHAT. (b) Proposed.*

In order to prove the theorem above, let us assume that a point inside,, takes the maximum value in the volume, i.e. $\tau_{kl}(\boldsymbol{x}_o) = (\tau_{kl}(\in \partial V)) = C_{maxV}$. Since there is a half-hyperboloid that goes through each point of the space, all the points besides $\boldsymbol{x}_o$ satisfying $\tau_{kl}(\boldsymbol{x}) = C_{maxV}$ will also take the maximum value. Therefore, all the points on the surface resulting from the intersection of the volume and the half-hyperboloid will take this maximum value, including those pertaining to the boundary surface ∂V. The existence of the minimum in ∂V is similarly deduced.

Following, a description of the proposed approach is given by analyzing a simple case where we want to estimate the location $\boldsymbol{x}_s$ of a sound source inside an anechoic space. In this simple case, the GCCs corresponding to each microphone pair are delta functions centered at the corresponding inter-microphone time-delays: $R_{mk\,ml}(\tau) = \delta(\tau - \tau_{kl}(\boldsymbol{x}_s))$. For example and without loss of generality, let us assume a set-up with $M = 3$ microphones, as depicted in Figure 2(a). Then, the source position would be that of the intersection of the three half-hyperboloids $\tau_{kl}(\boldsymbol{x}) = \tau_{kl}(\boldsymbol{x}_s)$, with $(k,l) \in \{(1,2), (1,3), (2,3)\}$. Consider now that, to localize the source, a spatial grid with resolution $r = 1$ m is used as shown in Figure 2(a). Unfortunately, the intersection does not coincide with any of the sampled positions, leading to an error in the localization task. Obviously, this problem would have been easier to solve with a two step localization approach, but the above example shows the limitations imposed by the selected spatial sampling in SRP-PHAT, even in optimal acoustic conditions. This is not the case of the approach followed to localize the source in Figure 2(b) where, using the same spatial grid, the GCCs have been integrated for each sampled position in a range that covers their volume of influence.

A darker gray color indicates a greater accumulated value and, therefore, the darkest area is being correctly identified as the one containing the true sound source location. This new modified functional is expressed as follows

$$P_n''(\boldsymbol{x}) = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \sum_{\tau=L_{kl1}(\boldsymbol{x})}^{L_{kl2}(\boldsymbol{x})} R_{m_k m_l}(\tau)\,. \qquad [7]$$

The problem is to determine correctly the limits $L_{kl1}(\boldsymbol{x})$ and $L_{kl2}(\boldsymbol{x})$, which are determined by the gradient of the IMTDF corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry, as explained in [14].

3.1 Computational cost

Let L be the DFT length of a frame and $Q = M(m-1)/2$ the number of microphone pairs. The computational cost off SRP-PHAT is given by [8]:

$$SRP\_PHAT_{cost} \approx 6.125Q^2 + 3.75QL\log_2 L + 15LQ(1.5Q - 1) + (45Q^2 - 30Q)v' \quad [8]$$

where v' is the average number of functional evaluations required to find the maximum of the SRP space. Since the cost added by the modified functional is negligible and the frequency-domain processing of our approach remains the same as the conventional SRP-PHAT algorithm, the above formula is valid for both approaches. Moreover, since the integration limits can be pre-computed before running the localization algorithm, the associated processing does not involve additional computation effort. However the advantage of the proposed methods relies on the reduced number of required functional evaluations for detecting the true source location, which results in an improved computational efficiency.

r	0.01	0.1	0.5
v'	$802 \cdot 10^5$	$802 \cdot 10^2$	641
SRP-PHAT	RMSE = 0.29	RMSE = 0.74	RMSE = 1.82
Proposed	RMSE = 0.21	RMSE = 0.29	RMSE = 0.31
SRC	RMSE = 0.34 (v' = 58307)		

Table 1. *RMSE for real-data experiment*

3.2 SSL comparative

Different experiments with real recordings were conducted to compare the performances of the conventional SRP-PHAT algorithm, the SRC algorithm and our proposed method. The set-up for these experiments consists on six omnidirectional microphones placed at the 4 corners and at the middle of the longest walls of a videoconferencing room with dimensions 5.7m x 6.7m x 2.1m and 12 seats. The measured reverberation time was T_{60}=0.28s. The resultant recordings were processed with 3 different spatial grid resolutions in the case of SRP-PHAT and the proposed method (r_1=0.01m,r_2=0.1m and r_3=0.5m). Note that the number of functional evaluations v' depends on the selected value of r, having $v'_1 = 802\ x\ 10^5$, $v'_2 = 802\ x\ 10^2$, and v'_3=641. The processing was carried out using a sampling rate of 44.1 kHz, with time windows of 4096 samples of length and 50% overlap, using continuous speech fragments obtained from the 12 seat locations. The average results in terms of Root Mean Squared Error (RMSE) are shown in Table 1, confirming that our proposed method performs robustly using a very coarse grid.

Note that, although similar accuracy to SRC is obtained, the number of functional evaluations is significantly reduced.

Figure 3 shows that, for a fine grid, there is no difference between traditional and modified SRP-PHAT method. Note that the GCC resulting from each pair of microphones crosses in the same point with equal accuracy. However, figures (a) and (b) of Fig. 4 show that the results of localization when a coarse grid is used in the GCC calculations have not equal accuracy if traditional or modified SRP-PHAT is applied. It can be seen that when a coarse grid is used in order to get lower computational cost, the traditional SRP-PHAT approach has not enough accuracy to find the SSL while the proposed modified SRP-PHAT is precise enough.

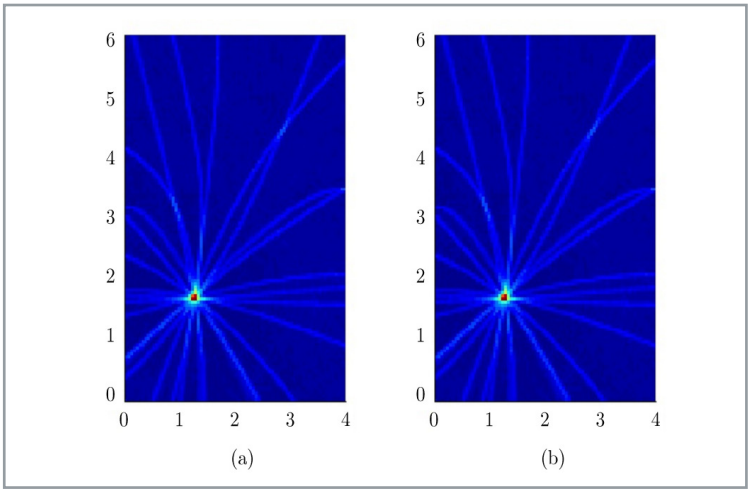

Figure 3. *Source likelihood map when a fine grid is used. (a) Traditional and (b) modified SRP-PHAT.*

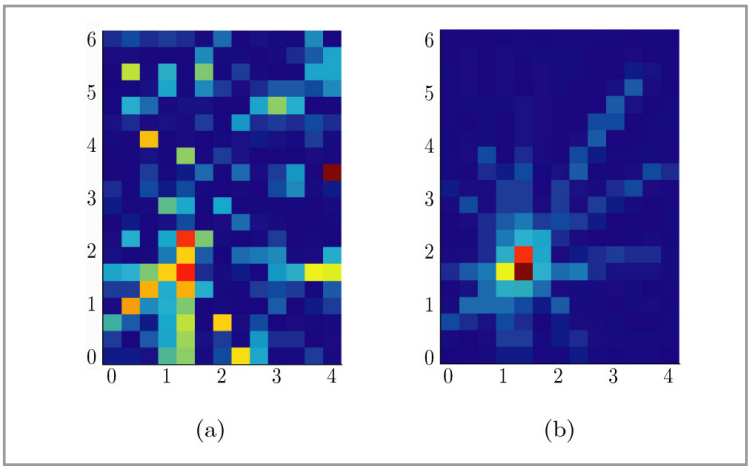

Figure 4. *Source likelihood map when a coarse grid is used. (a) Traditional and (b) modified SRP-PHAT.*

Method (T_{60})	Source 1 SNR = 10 dB			Source 2 SNR = 5 dB			Source 3 SNR = 0 dB		
r (m)	0.01	0.1	0.5	0.01	0.1	0.5	0.01	0.1	0.5
SRP (0.2 s)	100	90	79	99	89	63	89	71	35
Prop. (0.2 s)	100	100	100	100	99	99	90	89	87
SRP (0.7 s)	100	89	64	96	81	52	75	66	21
Prop. (0.7 s)	100	100	99	98	98	98	78	74	74

Table 2. *Performance in Terms of Percentage of Correct Frames.*

It is possible to discriminate between speech and non-speech frames by observing the peakedness of a set of accumulated estimates.

Another way to evaluate the benefits of our proposed approach is by looking at the results shown in Table 2, obtained by means of simulation. It shows the percentage of correct frames were the source was correctly located using our proposed approach and the conventional SRP-PHAT algorithm. Three source positions with different SNR were considered. A frame estimate is considered to be erroneous if its deviation from the true source location is higher than 0.4m, which is approximately the maximum deviation admissible for the coarser grid. Notice that, for the worst case (T_{60}=0.7 and SNR=0dB), the proposed approach is capable of localizing correctly the source with 74% correctness with r= 0.5m, which is approximately the performance achieved by the conventional algorithm using r=0.01m. Thus, our proposed approach provides similar performance with a reduction of five orders of magnitude in the required number of functional evaluations. Notice also that both methods perform almost the same in all situations when the finest grid is used.

4. Speaker detection and speech/non-speech discrimination

In the next subsections, we describe how active speakers are detected in our system, which requires a previous discrimination between speech and non-speech frames based on the distribution of location estimates. To this end, we model the probability density function of the obtained locations when there are active speakers and when silence and/or noise is present [15].

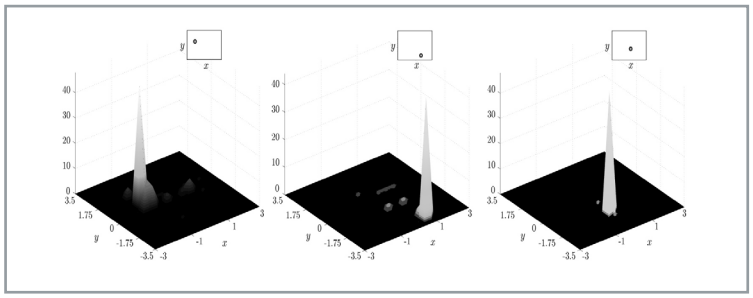

Figure 5. *Distribution obtained for three different speaker locations.*

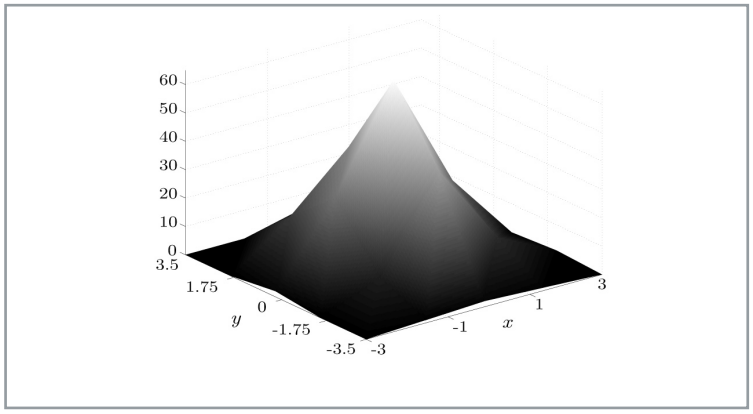

Figure 6. *Distribution for non-speech frames.*

4.1 Distribution of location estimates

The first step for speaker detection is to analyze the distribution of the location estimates $\hat{\boldsymbol{x}}_s$ when there is an active speaker talking inside a room from a static position. Figure 5 shows an example of three two-dimensional histograms obtained from different speaker locations in a rectangular room. It can be observed that, since the location algorithm is very robust, the resulting distributions when speakers are active are significantly peaky. Also, notice that the shape of the distribution is very similar in all cases but centered in the actual speaker location. As a result, we model the distribution of estimates as a bivariate Laplacian as follows:

$$p(\hat{\boldsymbol{x}}_s|H_s(\boldsymbol{x}_s)) = \frac{1}{2\sigma_x\sigma_y} exp^{-\sqrt{2}\left(\frac{|x-x_s|}{\sigma_x}+\frac{|y-y_s|}{\sigma_y}\right)},$$

[9]

where $p(\hat{\boldsymbol{x}}_s|H_s(\boldsymbol{x}_s))$ is the conditional probability function (pdf) of the location estimates under the hypothesis $H_s(\boldsymbol{x}_s)$ that there is an active speaker located at $\boldsymbol{x}_s=[x_s,y_s]$. Note that the variances σ_x^2 and σ_y^2 may depend on the specific microphone set-up and the selected processing parameters.

On the other hand, a similar analysis was performed to study how distribution changes when there are not active speakers, i.e. only noise frames are being processed. The resulting histogram can be observed in Figure 6, where it becomes apparent that the peakedness of this distribution is not as significant as the one obtained when there is an active source. Taking this into account, the distribution of non-speech frames is modeled as a bivariate Gaussian:

$$p(\hat{\boldsymbol{x}}_s|H_n) = \frac{1}{2\pi\sigma_{x_n}\sigma_{y_n}} exp^{-\left(\frac{x^2}{2\sigma_{x_n}^2}+\frac{y^2}{2\sigma_{y_n}^2}\right)},$$

[10]

where $p(\hat{\boldsymbol{x}}_s|H_s(\boldsymbol{x}_s))$ is the conditional pdf of the location estimates under the hypothesis that there are not active speakers, and the variances and are those obtained with noise-only frames.

4.2 Distribution speech/non-speech discrimination

In the above subsection, it has been shown that speech frames are characterized by a bivariate Laplacian probability density function. A similar analysis of location estimates when there are not active speakers results in a more Gaussian-like distribution, which is characterized by a shape less peaky than a Laplacian distribution. This property is used in our system to discriminate between speech and non-speech frames by observing the peakedness of a set of accumulated estimates:

$$\boldsymbol{C} = \begin{bmatrix} \hat{x}_s(n) & \hat{y}_s(n) \\ \hat{x}_s(n-1) & \hat{y}_s(n-1) \\ \dots & \dots \\ \hat{x}_s(n-N-1) & \hat{y}_s(n-N-1) \end{bmatrix} = [c_x\, c_y],$$

[11]

where N is the number of accumulated estimates in matrix $\boldsymbol{C}$. A peakedness criterion based on high-order statistics is evaluated. In probability theory and statistics kurtosis is the measure of the peakedness of the probability distribution of a real-valued random variable. Since the kurtosis of a normal distribution equals 3, we propose the following discrimination rules for active frames:

$$\mathrm{Kurt}(c_x)\begin{cases} \geq 3 \text{ speech} \\ < 3 \text{ non} - \text{speech} \end{cases}'$$

$$\mathrm{Kurt}(c_y)\begin{cases} \geq 3 \text{ speech} \\ < 3 \text{ non} - \text{speech} \end{cases}' \qquad [12]$$

where a frame is selected as speech if any of the above conditions is fulfilled.

5. Results

The new SSL method developed has been applied in a videoconference system where, by accurately estimating the various users physical locations, it would be possible to steer a video camera toward the currently active speaker.

To evaluate the performance of our proposed approach a set of recordings was carried out in a videoconferencing test room with dimensions 6.67m x 5.76m x 2.10m. A set of 6 omnidirectional microphones was placed on the walls of the room and 12 pre-defined target locations were used to select the active speaker seat (see Fig. 7). The experiment consisted in recording speakers talking from the different target positions (only one speaker at each time) with the corresponding space of silence between two talking interventions. The recordings were processed with the aim of evaluating the performance of our system in discriminating speech from non-speech frames and determining the active speaker so that the camera situated in the room can point at the correct seat. The discrimination between speech and non-speech frames was carried out by calculating the kurtosis of the last *N* estimated positions.

Table 3 shows the percentage of correctly detected speech (%SP) and non-speech (%N-SP) frames with different number of accumulated positions (N). Moreover, the processing was performed considering two different spatial grid sizes (0.3m and 0.5m). The percentage of speech frames with correct target positions (%T) is also shown in this table. It can be observed that, generally, the performance increases with a finer grid and with the number of accumulated estimates N. These results were expectable, since the involved statistics are better estimated with a higher number of location samples. Although it may seem that there are a significant number of speech frames that are not correctly discriminated, it should be notice that this is not a problem for the correct driving of the camera, since most of them are isolated frames inside speech fragments that do not make camera change its pointing target.

Grid res.	0.5 m				0.3 m			
N	5	10	15	20	5	10	15	20
% SP	52.5	60.4	70.0	74.0	68.9	70.7	83.1	85.4
% N-SP	75.9	64.8	70.9	72.7	81.4	70.9	81.5	82.3
% T	98.2				99.6			

Table 3. *Performance in Terms of Percentage of Correct Frames.*

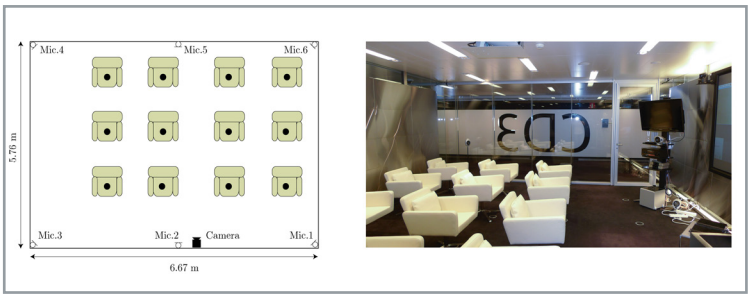

Figure 7. *Videoconferencing test room and microphones location (left) and real room (right).*

6. Conclusions

Sound source localization and speech/non-speech detection techniques have been presented in this work to be used in a multiparticipant videoconferencing environment with a microphone array system for a steering-camera application.

Based on the well known SRP-PHAT SSL method, a modified version of that technique that uses a new functional has been developed. The proposed functional is based on the accumulation of GCC values in a range that covers the volume surrounding each point of the defined spatial grid. The GCC integration limits are determined by the gradient of the inter-microphone time delay function corresponding to each microphone pair, thus, taking into account the spatial distribution of possible TDOAs resulting from a given array geometry. Our results showed that the proposed approach provides similar performance to the conventional SRP-PHAT algorithm in difficult environments with a reduction of five orders of magnitude in the required number of functional evaluations. This reduction has been shown to be sufficient for the development of real-time source localization applications.

In a videoconferencing environment where the sources are voices from different speakers, a speech/non-speech detection step is necessary to provide a robust steering camera system. For this reason the distribution of location estimates

has been obtained using the proposed SRP-PHAT functional. Our analysis shows that location estimates follow different distributions when speakers are active or mute. This fact allows us to discriminate between speech and non-speech frames under a common localization framework. The results of experiments conducted in a real room suggest that, using a moderately high number of accumulated location estimates, it is possible to discriminate with significant accuracy between speech and non-speech frames, which is sufficient to correctly detect an active speaker and point the camera towards his/her predefined location.

Acknowledgement

This work was supported by the Ministry of Education and Science under the project TEC2009-14414-C03-01.

References

[1] E. Ettinger and Y. Freund, "Coordinate-free calibration of an acoustically driven camera pointing system," in Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2008), Stanford, CA, 2008.

[2] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97), Washington, DC, 1997.

[3] H. Algahssi, "Eye Array Sound Source Localization". PhD thesis, University of British Columbia, 2008.

[4] N. Madhu, and R. Martin, "Advances in Digital Speech Transmission". Wiley, 2008, ch. Acoustic source localization with microphone arrays, pp. 135–166.

[5] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview". EURASIP Journal on Applied Signal Processing 2006 (2006), 1–19.

[6] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," IEEE Signal Process., vol. 12, no. 8, pp. 561–564, 2005.

[7] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in Microphone Arrays: Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag, 2001, pp. 157–180.

[8] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson, III, "Performance of real-time source-location estimators for a large-aperture microphone array," IEEE Trans. Speech Audio Process., vol. 13, pp. 593–606, 2005.

[9] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," EURASIP J. Appl. Signal Process., vol. 2003, nº. 4, pp. 338–347, 2003.

[10] J. H. DiBiase, "A high accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, Providence, RI, 2000.

[11] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array". In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007) (2007).

[12] H. Do, and H. F. Silverman, "A fast microphone array SRP-PHAT source location implementation using coarse-to-fine region contraction (CFRC)". In Proceedings of the IEEEWorkshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007) (2007).

[13] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," in IEEE Transactions on Acoustics, Speech and Signal Processing, SSP-24, 320–327, 1976.

[14] M. Cobos, A. Marti and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," in IEEE Signal Processing Letters, 18, no. 1, 71–74, 2011.

[15] A. Marti, M. Cobos and J. J. Lopez, "Realtime speaker localization and detection system for camera steering in multiparticipant videoconferencing environments", presented at 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), Prague, Czech Republic, 2011.

Biographies

Amparo Marti was born in Valencia, Spain, in 1983. She received the degree in Electrical Engineering from the Universitat Politècnica de València, Spain, in 2008 and the MSc. Degree in Telecommunication Technologies in 2010. Currently, she is a PhD grant holder from the Spanish Ministry of Science and Innovation under the FPI program and is pursuing her PhD degree in Electrical Engineering at the Institute of Telecommunications and Multimedia Applications (iTEAM) working in the field of multichannel audio signal processing.

Máximo Cobos was born in Alicante, Spain, in 1982. He received a telecomunications enegineer degree in 2006, an M.S. degree in telecommunications technologies in 2007 and the Ph.D degree in telecommunications in 2009, all of them from the Universitat Politècnica de València, Spain. His Ph.D dissertation on source separation for advanced spatial audio systems was awarded with the Ericsson Best Thesis Award on Multimedia Environments from the Spanish National Telecommunications Engineering Association (COIT). In 2009, he was a visiting researcher at the audio group of the Deutsche Telekom Laboratories in Berlin, where he worked in the field of audio signal processing for telecommunications. Currently, he is a postdoc researcher at the Institute of Telecommunications and Multimedia Applications in Valencia. His work is focused on the area of digital signal processing for audio and multimedia applications. He is interested in music and speech processing, microphone arrays, spatial audio and acoustics. Dr. Cobos is a member of the Audio Engineering Society (AES) and the IEEE and has published more than 40 papers in international journals and at renowned conferences.

Emanuel Aguilera was born in Buenos Aires, Argentina, in 1978. He received a telecommunications engineering degree from the Universitat Politècnica de València, Spain, in 2004. Currently, he combines his M. S. studies in computer science with his research at the Institute of Telecommunications and Multimedia Applications (iTEAM), where he has been working for 4 years on the area of digital signal processing for audio, multimedia and virtual reality. He is interested in wave-field synthesis, image processing, pattern recognition and real-time multimedia processing for telecommunications.

José Javier López was born in Valencia, Spain, in 1969. He received a telecommunications engineering degree in 1992 and a Ph.D. degree in 1999, both from the Universitat Politècnica de València, Spain. Since 1993 he has been involved in education and research at the Communications Department of the Universidad Politécnica de Valencia, where at present he is full professor. His current research activity is centered on digital audio processing in the areas of spatial audio, wave-field synthesis, physical modeling of acoustic spaces, efficient filtering structures for loudspeaker correction, sound source separation, and development of multimedia software in real time.

Dr. López has published more than 150 papers in international technical journals and at renowned conferences in the fields of audio and acoustics and has lead more than 25 research projects. He was workshop cochair at the 118th Convention of the Audio Engineering Society in Barcelona and has been serving on the committee of the AES Spanish Section for 8 years, at present as secretary of the section. He is a member of the AES, ASA, and IEEE.